REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT DATE	3. DATES COVERED *(From - To)*
02/28/2003	Final Technical	03/01/1999 - 02/28/2003

4. TITLE AND SUBTITLE

Multimodal Interaction for Wearable Augmented Reality Environments

5a. CONTRACT NUMBER

5b. GRANT NUMBER

N00014-99-1-0377

5c. PROGRAM ELEMENT NUMBER

01PRO7322-00

6. AUTHOR(S)

Philip R. Cohen

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Center for Human Computer Communication
OGI School of Science & Engineering, OHSU
20000 NW Walker Rd.
Beaverton, OR 97006

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research Regional Office
Seattle
1107 NE 45th Street Suite 350
Seattle, WA 98105-4631

10. SPONSOR/MONITOR'S ACRONYM(S)

ONR

11. SPONSORING/MONITORING AGENCY REPORT NUMBER

12. DISTRIBUTION AVAILABILITY STATEMENT

Approved for Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We describe an approach to natural 3D multimodal interaction in immersive environments. Our approach fuses symbolic and statistical information from a set of 3D gesture and speech agents, building in part on prior research on disambiguating the user's intent in 2D and 2.5D user interfaces. We present an experimental system architecture that embodies this approach, and provide examples from a preliminary 3D multimodal testbed to explore our ideas in augmented and virtual reality.

15. SUBJECT TERMS

Multimodal interaction, augmented reality, virtual environments, agent-based architecture, gesture recognition, input technique

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Philip R. Cohen
U	U	U		18	19b. TELEPONE NUMBER *(Include area code)* 503.748.1326

Final Report

Multimodal Interaction for Wearable Augmented Reality Environments

N00014-99-1-0377

Objective:

This effort is directed towards developing a new class of robust multimodal interaction technologies suitable for 3D augmented reality environments. It integrates speech and gesture recognition, natural language processing, and 3D virtual reality technology symbolically and statistically in an extensible, open architecture.

Accomplishments:
This project accomplished ground-breaking work in the integration of multiple modalities and knowledge sources into 3D augmented reality environments. Working with Columbia University, we extended the basic 2D architecture developed at OGI by incorporating 3D gesture recognition and information from reasoning about a scene into a full 3D multimodal augmented reality system. The system uses a detailed 3D model of the Columbia University Graphics Laboratory, including the physical objects within the lab, such as tables, chairs, walls, etc. The user's body is tracked using either a Flock of Birds magnetic tracker, or an Ascension IS900 wireless tracker. Extensions were made to the AAA multiagent architecture to enable it to support high-volume point-to-point communication. This work is detailed in the attached draft paper, which is being revised for publication.

Among the other accomplishments of this project were:

- Development of a 3-level recognition architecture (Members-Teams-Committee), which supports statistical and symbolic fusion for multimodal systems. This recognition architecture was shown to offer superior error handling (reducing error rates approx. 30%) over the existing symbolic-integration-only system. This recognition architecture was also deployed as a pen-based gesture recognizer for military symbols.
- Integration of the gesture recognizer into the first tangible multimodal system (Rasa). Rasa enables a military user to continue to employ his highly trained work style (using paper maps and Post-It notes), while the user's multimodal input is digitized simultaneously. Often, warfighters ignore the computers in their environment, preferring to employ paper-based tools. The Rasa system provides both the benefits of paper and digital systems. The system has been tested with members of the USMC and the US Army National Guard, and found to be preferred to paper alone, and to be robust to power and computer failures.

Publications:

Cohen, P.R., McGee, D., Oviatt, S.L., Wu, L., Clow, J., King, R., Julier, S., Rosenblum, L. Multimodal Interactions for 2D and 3D Environments, *IEEE Computer Graphics and Applications*, July/August 1999, pp.10-13.

Kaiser, E., and Cohen, P. R., Implementation testing of a hybrid symbolic/statistical multimodal architecture, to appear in Proceedings of the International Conference on Spoken Language Processing, Denver, September, 2002.

McGee, D. R., Cohen, P. R., Wu, L. (2000) "Something from nothing: Augmenting a paper-based work practice with multimodal interaction," in the Proceedings of the *Designing Augmented Reality Environments Conference (DARE'00)*, ACM Press: Copenhagen, Denmark, April 12-14, pp. 71-80.
http://www.cse.ogi.edu/CHCC/Papers/davePaper/dare00.final.pdf

McGee, D. R., Cohen, P. R. "Use what you've got: Steps toward opportunistic computing," to appear in the Proceedings of the *International Conference on Intelligent User Interfaces (IUI 2001)*, ACM Press: Santa Fe, NM, Jan. 2001. ftp://ftp.cse.ogi.edu/pub/tech-reports/2000/00-001-CHCC.pdf

McGee, D. R., Pavel, P., Adami, A., Wang, G. & Cohen, P. R.: "A Visual Modality for the Augmentation of Paper," in the Proceedings of the Workshop on Perceptive User Interfaces (*PUI'01*), ACM Press: Orlando, FL, Nov. 14-16, 2001.

McGee, D. R., Cohen, P. R., Wesson, R. M., & Horman, S.: "Comparing paper and tangible multimodal tools," in the Proceedings of the Conference on Human Factors in Computing Systems (CHI'02), ACM Press, (Minneapolis, MI, Apr. 20-25 2002).

Oviatt, Cohen, Wu, Vergo, Duncan, Suhm, Bers, Holzman, Winograd, Landay, Larson, & Ferro. "Designing the User Interface for Multimodal Speech and Pen-based Gesture Applications: State-of-the-Art Systems and Future Research Directions," to appear in *Human Computer Interaction* (to be reprinted in J. Carroll (ed.) *Human-Computer Interaction in the New Millennium*, Addison-Wesley Press: Boston).
http://www.cse.ogi.edu/CHCC/Publications/hci2000/hci.htm

Oviatt, S.L., "Taming Speech Recognition Errors Within a Multimodal Interface," in *Communications of the ACM*, September 2000.
http://www.cse.ogi.edu/CHCC/Publications/cacm9-2000/cacm9-2000.htm

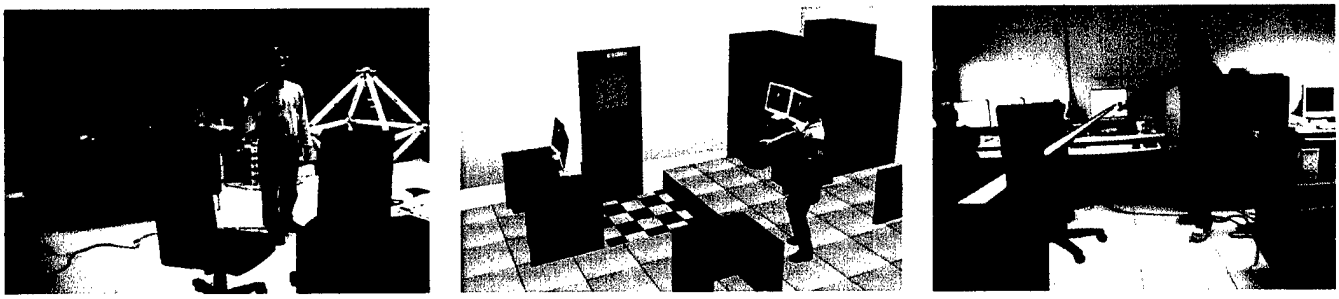
Oviatt, Sharon & Cohen, Philip. "Multimodal Interfaces That Process What Comes Naturally" *Communications of the ACM*, Vol. 43, No. 3, March, 2000, pp. 45-53.
http://www.cse.ogi.edu/CHCC/Publications/CACM3-2000.pdf

Wu, Lizhong, Oviatt, Sharon L., Cohen, Philip R., Multimodal Integration -- A Statistical View, *IEEE Transactions on Multimedia*, Vol. 1, No. 4, December 1999, pp. 334-341.

Wu, L. , Oviatt, S. L., and Cohen, P. R., From members to teams to committee: A robust approach to gestural and multimodal recognition, *IEEE Transactions on Neural Networks, Vol. 13, No. 4*, July 2002, pp. 1-11.

Patents Pending: "Augmenting and Not Replacing a Paper-Based Work Practice," D. McGee, P. R. Cohen, and L. Wu.

An Architecture for 3D Multimodal Interaction in Augmented and Virtual Reality



Abstract

We describe an approach to natural 3D multimodal interaction in immersive environments. Our approach fuses symbolic and statistical information from a set of 3D gesture and speech agents, building in part on prior research on disambiguating the user's intent in 2D and 2.5D user interfaces. We present an experimental system architecture that embodies this approach, and provide examples from a preliminary 3D multimodal testbed to explore our ideas in augmented and virtual reality.

CR Categories: H.5.1 (Multimedia Information Systems): Artificial, augmented, and virtual realities; H.5.2 (User Interfaces): Graphical user interfaces, natural language, voice I/O; I.2.7 (Natural Language Processing); I.2.11 (Distributed Artificial Intelligence): Multiagent systems; I.3.7 (Three-Dimensional Graphics and Realism): Virtual reality

Keywords: Multimodal interaction, augmented reality, virtual environments, agent-based architecture, gesture recognition, input technique

1 Introduction

Techniques for interacting in 3D worlds are usually derived from the direct manipulation metaphor—in order to perform an operation on something, you have to touch it. This style of interaction works well when the objects to be manipulated are known and at hand, and the means for selecting objects and other actions are relatively straightforward. Unfortunately, 3D interaction often breaks all these rules—the objects of interest may be unknown and may reside far away in the 3D environment, and there may be far more possible actions that can be performed than GUI function buttons or menu items can realistically provide. To cope with these problems, some researchers have taken the direct manipulation style of interaction to extremes, creating devices with many buttons and modes [Fröhlich et al. 2000], arbitrarily stretchable "arms" [Poupyrev et al. 1996], and 3D menus [Liang and Green 1994].

We contend that sometimes too much functionality has been forced on too impoverished a communications channel (3D arm/hand motions), and that by incorporating multimodal interaction, the burden of various interactive functions can be off-loaded to appropriate modalities, such as speech and gesture, in a synergistic fashion. In particular, by incorporating speech into the interface, the user could describe unseen objects and locations or invoke functions, while her hands and eyes may be engaged in some other task.

Multimodal interface architectures need to cope first and foremost with uncertainty. Recognizers return a set of classification hypotheses, each of which may be assigned a score, such as a posterior probability. Moreover, language is ambiguous, and thus even a correctly recognized utterance can lead to multiple meaning hypotheses. Likewise, trackers have errors, gestures are uncertain and their meanings are ambiguous, and even a correct gesture (e.g., selection) can have multiple interpretations (e.g., what is being selected). Given all this uncertainty, it is perhaps surprising that few if any multimodal systems that support speech with 3D gestures in 3D virtual reality (VR) or augmented reality (AR) environments are able to deal directly with that uncertainty.

To address these issues, we present an experimental architecture for 3D multimodal interaction with real and virtual objects and show how it can reduce errors by fusing symbolic and statistical information derived from speech, gesture, and the environment. We begin by describing related work in Section 2. Then, we introduce an application scenario that we are exploring in Section 3, including examples from a live run of a testbed implementation of our architecture. We describe the architecture itself in Section 4, and present our conclusions and future work in Section 5.

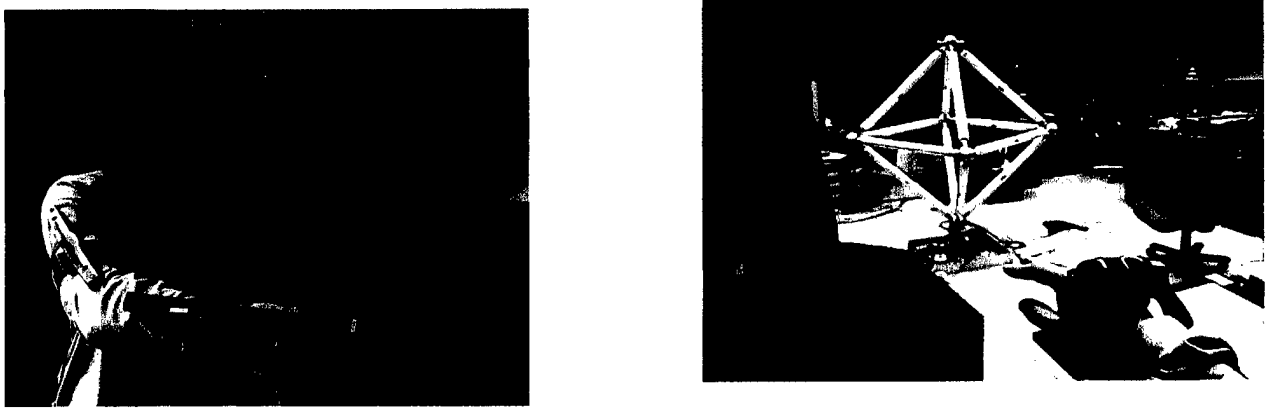


Figure 1. (a) User with 6DOF trackers on see-through, head-worn display, upper arm, lower arm, and hand. (b) AR user's view (imaged through a tracked video camera), with virtual desk and chair on the left.

2 Related Work

We can categorize research in multimodal interaction along two axes—the dimensionality of the gestures (2D or 3D), and the dimensionality of the environment (2D, 2.5D, or 3D).

Overall, most multimodal research adopts some version of a *late-fusion architecture* [Oviatt et al. 2000], in which multimodal integrators fuse meaning structures subsequent to recognition. Yet, relatively few projects consider the issues involved in the management of uncertainty across modalities.

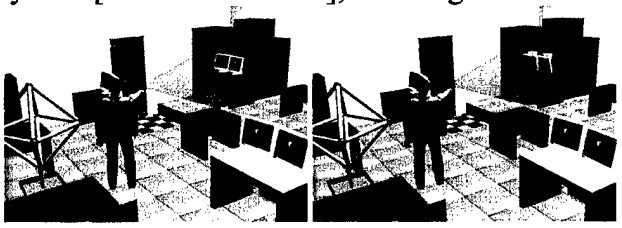
2.1 Multimodal Interaction with 2D Gestures for 2D or 2.5D Environments

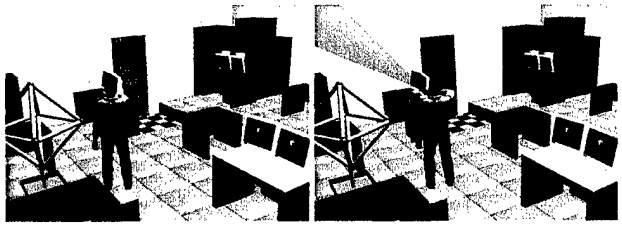
Many researchers have investigated multimodal 2D map-based interactions, dating back at least as far as the Cubricon system [Neal and Shapiro 1991]. That system, and others, such as Chorus [Tyler et al. 1991], Eucalyptus [Wauchope 1994], and Shoptalk [Cohen et al. 1989], demonstrated that the natural language subsystem could choose the correct referent among a set of selected items, based on the semantics of the corresponding deictic noun phrase that was uttered or typed. Unfortunately, the architectures built for those early systems placed the natural language parser in charge of multimodal interpretation; thus, if a deictic noun phrase was not uttered, no gesture would be interpreted. Worse still, the gesture referents had little influence on the ultimately chosen multimodal interpretation. Moreover, these systems were insensitive to uncertainties in their inputs. Landragin [2002] has recently

proposed 2D algorithms that could be used to fuse language, pen-based gesture, and object identification, suggesting the use of Gestalt principles to discover objects when none meet the linguistic criteria. Lacking an implementation however, his algorithms have not been tested with real-world errorful data

QuickSet is a 2D multimodal pen/voice map-based tool that enables users to create scenarios by speaking and sketching [Cohen et al. 1997]. QuickSet offers a late-fusion architecture, and integrates modalities semantically via unification of *typed feature structures* (directed and typed attribute-value graphs that can contain logical variables) that represent the meanings of the inputs [Johnston et al. 1997]. Strictly using semantic information, this system has been shown to offer *mutual disambiguation of modalities* [McGee et al. 1998, Oviatt, 1999], resulting in error rate reductions of 19–40% [Oviatt 1999, 2000] as compared to the performance of the unimodal recognizers. (Mutual disambiguation of speech and gesture occurs when one or the other or both of the top-scoring speech candidate and the top-scoring gesture candidate fails to participate in the top-scoring multimodal result.) Recent versions of the system employ a hybrid statistical/semantic integration architecture [Wu et al. 1999; Kaiser and Cohen 2002], which enables the system to weight the contributions of each modality based on the input, the relative reliabilities of the recognizers, and the overall statistical properties of the domain. The weighting parameters are learned from a labeled corpus of representative interactions.

The QuickSet architecture has been integrated into the Naval Research Laboratory's Dragon VR system [Cohen et al. 1999], resulting in a multimodal system that employs 2D gestures in a 2.5D world of topographical maps and aerial images—digital "ink" drawn on a 2.5D topographical map is projected onto the surface of a 3D scene. This system inherits the mutual disambiguation capabilities discussed, and thus allows speech to overcome gesture recognition errors, and vice versa; however, gestures are still limited to 2D.





2.2 Multimodal Interaction with 3D Gestures for 2D Environments

Multimodal 3D interaction that includes speech dates back at least to Bolt's pioneering Put-That-There system [Bolt 1980], in which speech was integrated with 3D magnetic tracking of a user's arm in order to manipulate a 2D world. Motivated by Bolt's landmark work, numerous researchers have investigated multimodal 3D interaction for 2D worlds. Koons et al. [1993] present a system that tracks 3D hand-based pointing gestures, speech, and gaze, and discuss its extension to other kinds of 3D gestures. The system copes with linguistic and referential ambiguity, but not erroneous recognizer inputs. Lucente et al. [1998] describe a system using IBM's ViaVoice speech recognizer and a vision-based hand and body tracker that enables a user to manipulate large objects on a display screen. Because of the size of the objects, it does not appear that reference resolution and uncertainty was of particular concern, nor was any error correction capability discussed. Similarly, Poddar et al. [1998] discuss a sophisticated system that understands speech and natural 3D gesture in a 2D environment, in which the speech and gesture of cable television Weather Channel narrators are analyzed as they described the movement of weather fronts across a map.

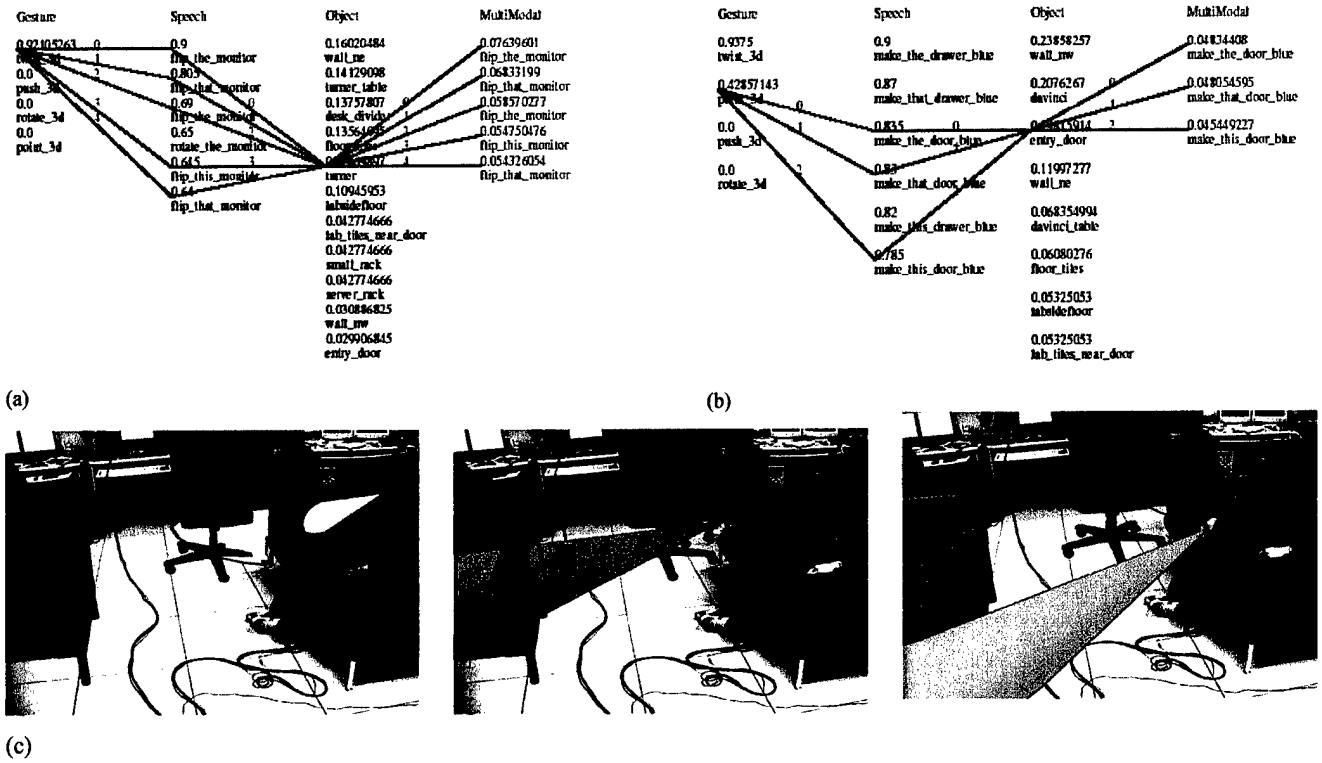


Figure 2. Three sequences of interactions in the environment. (a) "Make the door blue": Before and after VR images, with parallel coordinates plot showing multimodal disambiguation (see Section 4.3). (b) "Flip the monitor": Before and after VR images, with parallel coordinates plot showing multimodal disambiguation (see Section 4.3). (c) AR images showing conical region of interest interacting with real (couch and floor) and virtual (chairs) objects in the environment, as user selects chair on the left side of images and changes its color.

2.3 Multimodal Interaction with 3D Gestures for 3D Environments

It is becoming clear that users would like to have the ability to engage in multimodal virtual reality interaction. Hauptmann [1989] reported that in a Wizard-of-Oz (WOZ) study of spoken and two-handed interaction, 58% of subjects clearly preferred use of both modalities to any single one (19% gestures or 22% speech). Similarly, in a WOZ study of speech and single-handed VR interaction, Anonymous [in press] found that given the freedom to use speech, gesture, or multimodal interaction, 60% of the subjects habitually used multimodal interaction for navigation and manipulation, 30% habitually used speech only, and 10% habitually used gesture. (Habitual use meant use during more than 60% of the possible interactions.)
Many initial steps were taken that motivated building multimodal VR systems [Weimer and Ganapathy 1989; Bolt and Herranz 1992]. More recently, Billinghurst et al. [1995] discuss a multimodal virtual surgery assistant that incorporates a two-gesture vocabulary (point and grasp), simple speech recognition (vocabulary and grammar unspecified), speech synthesis, and an expert-system surgical assistant subsystem. Based on user queries, the system can overlay graphics on anatomical images and illustrate paths of surgical approach. Duncan et al. (in [Oviatt et al. 2000]) present a multimodal 3D virtual aircraft maintenance assistant that includes an avatar driven by the user's tracked limbs, gesture recognition (seven Cyber-glove–based gestures), spoken natural language input, and semantic fusion of temporally co-occurring input modes. LaViola [2000] describes a multimodal 3D system incorporating two-handed PinchGlove gestures and a 20-word spoken vocabulary. He uses a "show-and-ask" paradigm, in which tools are created at the 3D location at which a user's virtual hand is located. Users report that the multimodal interaction style is appealing, but spoken errors in open-microphone demonstration situations (with

ongoing conversations with other people) undermined usability. Krum et al. [2002] report on a multimodal VR system that uses finger gestures and speech to support 3D navigation. Althoff et al. [2001] discuss a system that employs techniques similar to [Johnston et al., 1997] for multimodal VR navigation. They suggest the use of genetic algorithms to address statistical fusion of information, but no implementation or results are discussed. The work most comparable to ours is that of Latoschik [2002], who developed a system based on augmented transition networks (a natural language processing technique used in the 1970s–80s). The system indeed merges speech and gesture, but no mention is made of handling of recognition errors and the possibility of mutual disambiguation. Other researchers have investigated multimodal interfaces for perceptive environments [Brummit et al. 2000] and robot control [Bauckhage et al. 2002, Iba et al. 2002, Perzanowski et al. 2000].

In most of these 3D systems, specific 3D gestures have been designed for each individual application, which users needed to learn to perform properly. Other work has attempted to analyze people's natural gestures [Quek et al. 2001; Anonymous in press] and to develop recognizers for them (e.g., [Kettebekov et al. 2002]), resulting in a greater naturalness, but also greater likelihood of errors because of increased variability. In summary, with the exception of QuickSet, none of the above systems is organized to manage uncertainty and the attendant recognition errors, and thereby offer (mutual) disambiguation of modalities. In this paper, we discuss how an architecture similar to that used in QuickSet for 2D gestures and 2D–2.5D environments can be extended to handle 3D gestures and to take uncertainty into account in immersive 3D VR and AR environments.

3 Application Scenario

We illustrate the kinds of interactions that we address with an example of manipulating real and virtual objects in a simple interior decoration scenario. The user is standing in the room, with three six-degree-of-freedom (6DOF) trackers attached to the right hand, right wrist, and right upper arm, as illustrated in Figure 1(a). The user also wears a see-through head-worn display, which has a fourth tracker attached to it, so that the head can be tracked, allowing graphics to be overlaid correctly onto the surroundings. The user's view is shown in Figure 1(b). Figure 2 shows both fully synthesized VR images and videomixed AR images that combine real and virtual objects (Section 4.1.1). In Figure 2(a–b), the VR user uses speech and hand gestures to flip a dual monitor configuration and change the door's color to blue. (The graphs shown in the image demonstrate the multimodal disambiguation process described in Section 4.3.) In Figure 2(c), the AR user sweeps a bounding volume attached to their hand, which interacts with real objects (the couch and floor) and virtual objects (the two chairs) as they change the color of the leftmost chair.

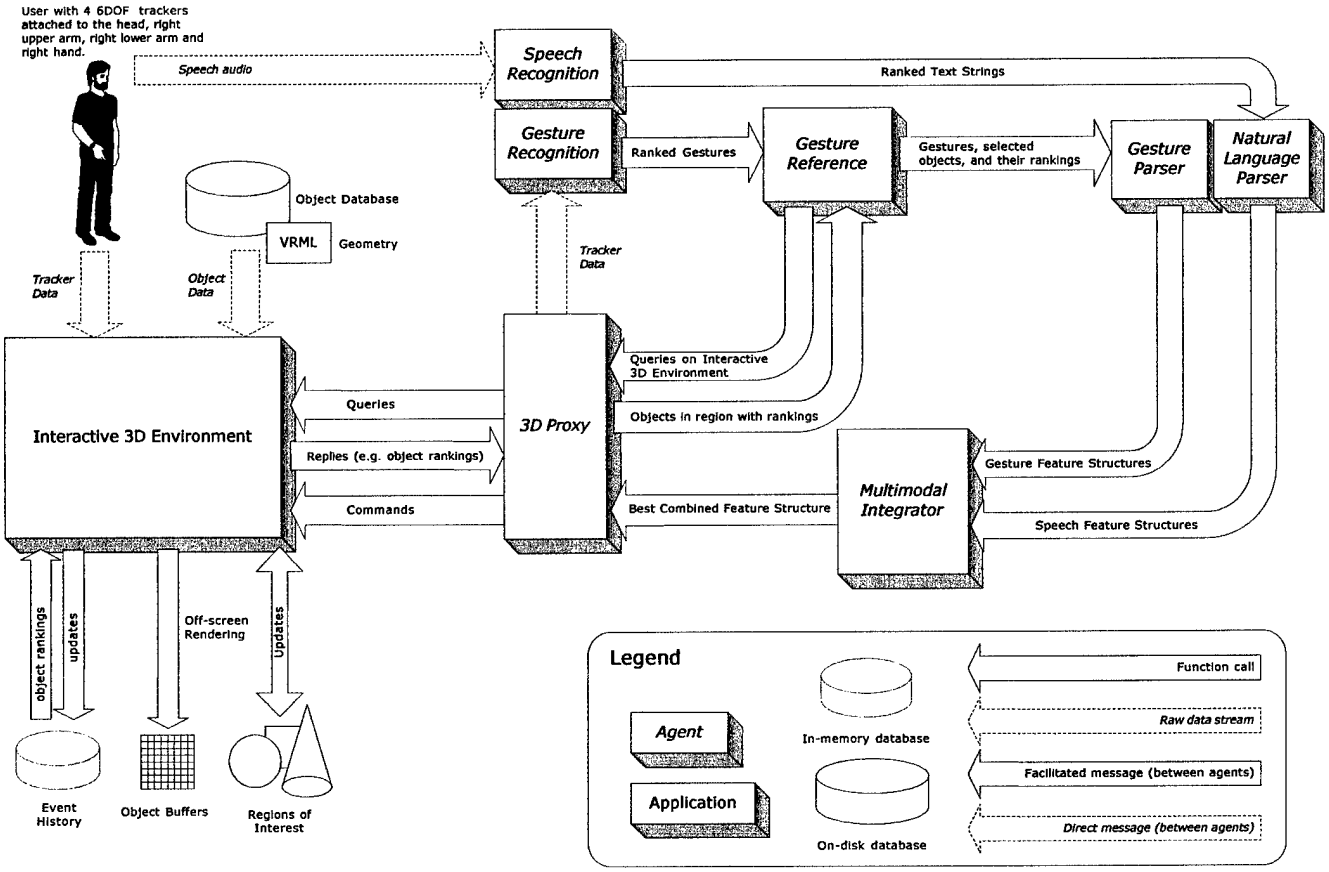


Figure 3. Multimodal interaction architecture.

4 System Architecture

The multimodal recognition architecture that we have developed consists of six components: an agent communication infrastructure, an interactive 3D environment (and its 3D Proxy agent), a gesture reference agent, a set of unimodal recognizer agents (one for each individual modality: speech and 3D gesture), a set of parser agents (one per modality), and a multimodal integrator agent. Their interactions are shown in Figure 3. The agent communication infrastructure [Anonymous 2000], implemented in Prolog and Java, is the underlying distributed communication system that connects all other components, supporting both facilitated communication (through a blackboard) and direct peer-to-peer communication. The interactive 3D environment is responsible for capturing raw user interactions, handling virtual world state changes, visualizing the interaction as VR or AR, and performing geometric processing needed to determine candidate referents for manipulation; it communicates with the rest of the components through its 3D Proxy agent. The gesture reference agent maintains the relationship between sensors and body positions, and evaluates the results of 3D gesture recognition to request the list of objects that were captured by the tracker regions of interest from the 3D proxy agent. The unimodal recognizers generate lists of recognition hypotheses, each with an associated score, and pass these lists to their parsers (in the case of gesture, mediated by the gesture reference agent). Each parser translates the lists to meaning fragments, delineates the potential ambiguities of actions in each mode, and forwards these fragments to the multimodal integrator for fusion. The top-scoring fused command meaning is then sent to the interactive 3D environment as an update.

4.1 Interactive 3D Environment

The interactive 3D environment supports manipulation of domain objects (e.g., by 3D geometric transformations and appearance changes), visualization for AR and VR (including effects such as highlighting), and head and body tracking using several different 6DOF trackers (InterSense IS-900 and Ascension Flock of Birds). It is the only component not directly implemented within the agent infrastructure, and therefore requires the 3D proxy agent to communicate with the rest of the system. It uses Java3D and runs on a dual Athlon MP 2.0 computer, with 1GB RAM and an NVIDIA Quadro4 750 graphics card. In the examples shown in this paper, our environment is represented by a 3D model of one of our laboratories, in which each object is tagged as either "real" (corresponding to a physical object) or "virtual."

4.1.1 Rendering for Augmented and Virtual Reality

Real and virtual objects are rendered with complete material appearance properties for VR displays; however, for AR, the real objects are instead rendered in a designated color, with no shading or lighting. We render real objects in black (Java3D does not support selectively disabling the frame buffer) when using a tracked optical see-through head-worn display (Sony Glasstron LDI-D100B), allowing virtual objects to occlude and be occluded by real objects. To create the AR images and video in this paper, we render the real objects in a unique key color [Azuma 1997], and use a video switcher (Videonics MX Pro) to chromakey the frame buffer with the video stream from a 6DOF-tracked

NTSC camera (Costar DSP 21X) whose tracker is used to control the virtual camera.

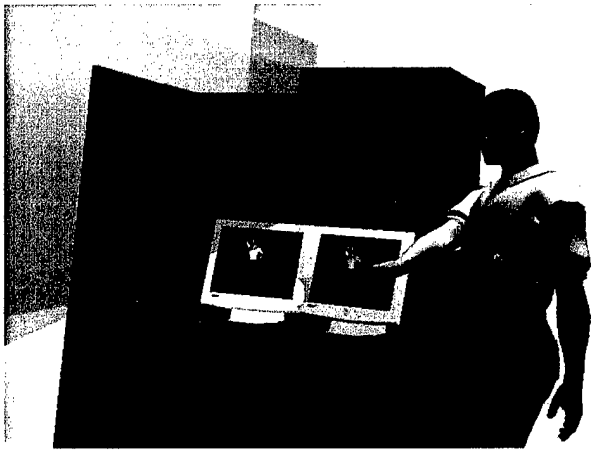


Figure 4. VR avatar controlled by tracked user, showing attached regions of interest.

4.1.2 *Regions of Interest and Object Ranking*

An important task of the interactive 3D environment component is to find geometric correlates for the semantic meaning of terms such as "that," "here," and "there," as well as to facilitate selection of objects. We accomplish this through *regions of interest,* volumes controlled by the user as she interacts with the environment. These regions are used to select and manipulate objects in the scene. Our current region of interest implementation includes four primitives: cuboids, cylinders, cones, and spheres. For example, Figure 4 shows two cones emanating from the user's eyes to approximate the field of view, a sphere around the user's hand to represent a volume that would encompass objects that are nearby and within reach, and another cone emanating from the user's hand, representing a volume that is intersected with potential pointing targets. The regions of interest are tested at each frame for intersection with objects in the environment to determine whether or not an object is within the region. The multimodal agents query the environment for objects and their relation to a region as needed.

The environment has the potential to contribute significantly to multimodal fusion. In particular, the regions of interest are used to provide a ranking for each object within them, based on the likelihood of that object being one that is intended for selection, just as gesture and speech candidates are ranked.

We currently provide four different types of region-of-interest–based rankings for an object: time, stability, visibility, and center-proximity. These rankings are relative to a specific object's behavior in a certain region of interest during a time period specified by the gesture reference agent.

The *time* ranking of an object is derived from the fraction of time the object spent in a region over a specific time period: the more time the object is in the region, the higher the ranking. The time ranking (T_{rank}) is defined as

$$T_{rank} = \frac{T_{object}}{T_{period}}, \qquad 1 \geq T_{rank} > 0,$$

where T_{object} is the total time an object is present in the region during T_{period}, which is the length of the time period of interest.

The *stability* ranking of an object expresses the stability of the object's presence in the region relative to other objects. We currently calculate this based on the number of times the object and other objects went in or out of the region during the time period. The fewer times an object enters or exits the region, the more stable we consider it. The object(s) with the least entries/exits are ranked highest, and the object(s) with the most entries/exits are ranked lowest. The stability ranking (S_{rank}) is defined as

$$S_{rank} = \frac{\max(E_{all\ objects}) + 1 - E_{object}}{\max(E_{all\ objects}) + 1}, \qquad 1 \geq S_{rank} > 0,$$

where $\max(E_{all\ objects})$ is the most times any object passed into or out of the region, and E_{object} is the number of times the particular object passed into or out of the region. (The most stable objects never leave the region, and thus have $E_{object} = 0$ and $S_{rank} = 1$.)

The *visibility* and *center-proximity* rankings of an object reflect its visibility relative to selected regions of interest. We compute the visibility of a conical region of interest by rendering into an off-screen object buffer [Atherton 1981] a low-resolution version of the scene from a center of projection at the cone's apex, cropped to the cone's cross-section. Each object is rendered with a unique color, allowing it to be identified in the frame solely by its pixel color, as shown in Figure 5.

We currently generate two such object buffers, one for an eye cone and one for the hand cone. The visibility ranking (V_{rank}) is defined as

$$V_{rank} = \frac{\sum visiblePixels_{object}}{\sum pixelsInFrame}, \qquad 1 \geq V_{rank} \geq 0,$$

where $visiblePixels_{object}$ are the visible pixels an object has in a frame, and *pixelsInFrame* are all the pixels in the frame.

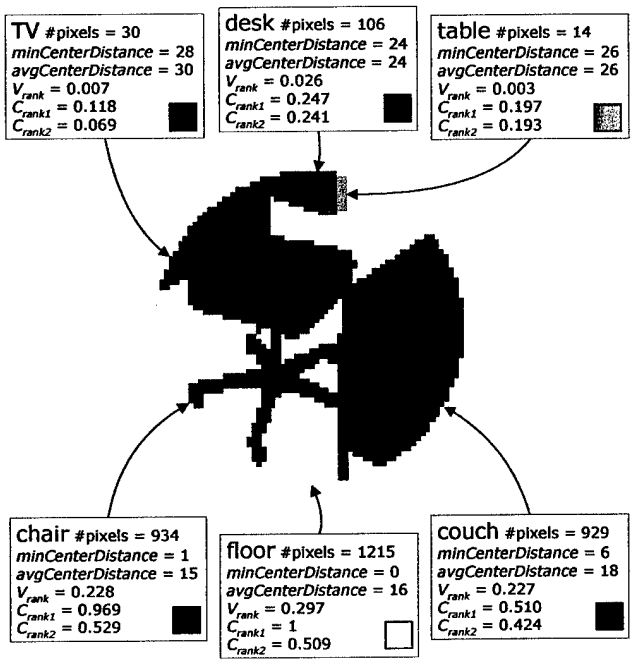


Figure 5. Off-screen object buffer (64 × 64 pixel rectangle). Six objects (a chair, a couch, the floor, a TV, a desk and a table) found in the object buffer. Each object is listed with visibility and center-proximity relevant information, including the ranking.

The center-proximity rankings indicate how close the visible portion of an object was to the center of the region. The distance is calculated as the Euclidean distance from the center of the object buffer. The center-proximity rankings (C_{rank1} and C_{rank2}) are defined as

$$C_{rank1} = 1 - \frac{minDistanceToCenter_{object}}{maxDistanceToCenter_{frame}}, \qquad 1 \geq C_{rank1} \geq 0,$$

and

$$C_{rank2} = 1 - \frac{avgDistanceToCenter_{object}}{maxDistanceToCenter_{frame}}, \qquad 1 \geq C_{rank2} \geq 0,$$

where $maxDistanceToCenter_{frame}$ is the maximum Euclidean distance any pixel can have to the center of the object buffer, $minDistanceToCenter_{object}$ is the smallest Euclidean distance of any pixel of the specific object, to the center of the object buffer, and $avgDistanceToCenter_{object}$ is the average Euclidean distance from the center, based on all visible pixels for the specific object.
Included with the object rankings are the time the object spent in the region (T_{object}) and the number of times the object went in/out of the volume (E_{object}). These absolute values allow thresholds to be used when classifying objects.

4.1.3 Event History

The event history uses an in-memory database to store information about all objects in the regions of interest, and supports complex queries through SQL (Structured Query Language). For example, queries are sent from the gesture reference agent to aid disambiguation, such as (paraphrased from SQL):

"Which objects were in region R between time T(a) and T(b) and what are their object rankings?"

The visibility and center-proximity rankings are computed and stored each time the object buffers are rendered. On retrieval, we currently compute their average values for the specified time period. The time ranking and stability ranking are computed when the query is executed.

4.2 Unimodal Recognition and Parsing

The source of all interaction in our system is the user's speech and the tracker data that represents her motion. These unimodal data streams are processed independently and in parallel, and then fused in the multimodal integrator agent.

4.2.1 Natural Language

Our speech agent uses an off-the-shelf recognition engine—the Microsoft Speech API 4-compatible *Dragon Naturally Speaking 6.* Results from the speech recognition engine are passed to the natural language parser as a list of probability-ranked, time-stamped text strings. The parser interprets raw text strings such as "Move that couch there," generating a potentially ambiguous set of meaning representations embodied in typed feature structures.

4.2.2 3D Gesture

Our 3D hand-arm gesture recognition agent receives and analyzes tracker data and sends messages to the gesture reference agent and gesture parser agent whenever supported 3D gestures are encountered in the tracker data stream. We consider the tracker data stream for a particular sensor to be in a *stationary state* whenever the sensor's reports do not vary over time by more than a tunable offset. By detecting stationary states, the recognizer determines explicit start and end points for classification without the need for specific user-defined positions or trigger mechanisms for locating start/end gesture points.

Recognition is based on a model of the body for which we track human movements and a set of rules for those movements. These rules were derived from an evaluation of characteristic patterns we identified after analyzing sensor profiles of the movements underlying the various gestures. Currently, the gesture recognition agent supports four kinds of gestures: pointing, twisting the hand about the index finger, rotating the hand about the wrist, and pushing with the palm up or down.

4.2.2.1 Pointing

Based on empirical study and analysis of collected gesture data [Anonymous 2002], we characterize a pointing gesture as:

1. *A hand movement between two stationary states,*
2. *Lasting between 0.5 and 2.4 seconds,*
3. *Whose dynamic patterns are characterized by a smooth and steady increase/decrease in the spatial component values,*
4. *Whose head direction and pointing direction form an angle below a heuristically determined threshold, and*
5. *Whose pointing direction and the imaginary direction determined by the upper arm forms an angle below some certain threshold.*

The fourth condition implicitly assumes that head orientation is a reliable indicator of the direction of the user's attention. Estimating where a person is looking based solely on her head orientation is a plausible simplification used to determine the focus of attention of the user without having to perform eye gaze tracking [Stiefelhagen 2002]. In VR usage of the system, standing before a wall screen without the use of a head mounted display, we use neither 3, the smoothness constraint, nor 4, the head/point direction condition. Instead we derive probabilities from normalized accumulators over the change in relative angles and positions between the hand and wrist sensors. A similar rule-based analysis of hand twisting and hand rotating can be given using the quaternion components provided by the sensor. A more extensive experimental analysis of natural gesture in virtual reality is being undertaken based on a "Wizard of Oz" study [Anonymous in press].

4.2.2.2 Twisting, Rotation and Pushing

Twisting the hand palm-down to palm-up about the index finger, rotating the hand side-to-side about the wrist, and waving or pushing with the hand up-and-down are similar rotational movements, occurring about two orthogonal axes. To recognize such gestures, we analyze the hand rotation information and changes in wrist/hand relative angles using the quaternion components provided by the respective sensors. We characterize a hand twisting about the pointing direction (palm-down to palm-up) as no or little change in the relative hand/wrist angles while wrist position is relatively stable and the direction the palm faces in changes significantly. A side-to-side hand rotating gesture about the stationary wrist is characterized by little change in the face direction of the palm with significant change in where the side-to-side direction the fingers are oriented. The pushing gesture is like a wave. It is characterized by wrist stability with little side-to-side change in the direction of the fingers while at the same time there is significant change in the up-down direction of the fingers.

4.3 Multimodal Integration

The multimodal integrator agent determines which combinations of speech and gesture interpretations can produce an actual command, using the approach of [Johnston 1998]. The basic principle is that of typed feature structure unification [Johnston et al. 1997], which is derived from term unification in logic programming languages. Here, unification rules out inconsistent information, while fusing redundant and complementary information through binding of logical variables that are values of "matching" attributes. The matching process also depends on a type hierarchy. A set of multimodal grammar rules specify, for a given task, which of these speech and gesture interpretations should be unified to produce a command. For example, a rule that unifies a pointing gesture interpretation with a spoken language interpretation might specify that a 3D pointing gesture selecting an office object could be unified with speech referring to that same type of office object.

Consider the following examples of multimodal integration derived from a sample run of our testbed, in which candidates for speech, gesture, and object selection are ranked from best to worst. Figure 2 includes two parallel coordinate plots [Inselberg and Dimsdale 1990], which are generated by our testbed automatically when a multimodal command succeeds. In these plots, each vertical axis represents the *n*-best list of the interpretation results of each modality (gesture and speech) and of the object selection rankings of the interactive 3D environment. Each item on the axis for that interpretation result is displayed with its probability and semantic information. The blue and red polylines represent possible final commands made up by unifying these components, where the red polyline is the best final command, which was actually executed. When the red polyline dips below the top entry on any axis, the architecture has chosen an interpretation of the input that was not the highest-ranked interpretation for its modality.

In the first case shown in Figure 2(a), the speech disambiguates the object selection, making the top command the one that flips the monitor (of the computer "turner"), not the wall. The monitor is well down on the object list, but it becomes part of the final command by virtue of the disambiguation provided by the speech.

In the second case in Figure 2(b), the multimodal architecture employs mutual disambiguation to resolve a very sloppy user command. The user points in the general direction of the door (entry_door), but the cone includes the monitor (partially blocked by the user's head), the door, and the wall in the object selection. In addition, the user issues an ambiguous speech command, "Make the door blue," in which he mispronounces "door" as "drawer." Finally, the user doesn't point very precisely, and the top gesture interpretation is a twist. The top-ranked speech, the top-ranked gesture, and the top-ranked object are all wrong. However, the speech, gesture, and object all disambiguate one another, and this mutual disambiguation, made possible by the multimodal architecture, resolves the command despite the errors. The top combination—the one executed as a command—is the correct one reflecting the user's intent

These examples of mutual disambiguation of multimodal inputs

employ information about the environment to disambiguate both speech and gesture. The need for such processing will only increase as each of the modalities scales up in complexity and variability (e.g., larger vocabulary and grammar, freer use of gestures, and more complex scenes).

5 Conclusions and Future Work

The architecture described here improves upon the current state-of-the-art in 3D multimodal research by reducing uncertainty and ambiguity through the fusion of information from a variety of sources. While our architecture is similar to what has been reported for 2D and 2.5D interaction, it takes advantage of the 3D environment, which provides additional sources of information (e.g., object identification, head tracking, and visibility). The system is designed to uncover the "best" joint interpretation of speech, gesture, and object identification given semantic and statistical properties. We have shown examples in which each modality compensates for errors in another. In an early pilot test, the modality disambiguation capabilities reduced the multimodal error rate by 16.7%. (Note that relative error rate reductions are the standard way to report recognizer improvements in the spoken language community.)

Our work also departs from current practice by recognizing 3D selection as but one possible interpretation of a gesture. Rather than require devices with buttons and modes, the system attempts to employ recognition technologies that perform selection as a byproduct of determining the best interpretation of the multimodal inputs. Such a model is particularly useful for applications in which displaying the regions of interest, or cursors, is inappropriate. After all, people normally interact with each other with only a general notion of where their interlocutors are pointing. Because of its distributed, multi-agent architecture, our system is readily extendable to collaborative multi-user interaction.

We are currently planning a number of improvements to our initial implementation:

- *More natural gesture recognition.* Based on data collected during a multimodal WOZ VR experiment [Anonymous in press], natural gestures are being identified, classified by hand, and then provided as a corpus for training hidden Markov model-based gesture recognizers. We believe that this style of gesture recognizer will improve upon the rule-based recognizer discussed here.
- *Learning the utility of recognition features.* Object identification provides a number of parameters whose importance is currently unknown. Given the hidden Markov model-based gesture recognizers discussed above, we will collect a corpus of user interactions and their associated parameters. The collection of features will be provided as input to a hierarchical statistical classifier (e.g., based on leveled HMMs [Oliver et al. 2002] or the approach of [Wu et al. 1999]), which will assign weights to speech, gesture, and object features.
- *A more comprehensive vocabulary and grammar.* While such vocabulary and grammar extensions can increase

expressive power, they can also result in more speech recognition errors and linguistic ambiguities. We anticipate that our system will be able to cope more gracefully than others in the literature, given the error rate reductions observed in current 2D systems that perform mutual disambiguation.

We have described how a multimodal architecture can make it possible to interact with immersive 3D AR and VR environments in a more natural fashion than is possible in previous approaches. To validate our hypothesis, we designed and implemented a testbed based on this architecture, which has been used to make the figures included in this paper. Based on user studies that we will perform within our testbed, we expect to significantly improve the algorithms currently used to generate statistics for the interactive 3D environment selection mechanism and the gesture recognition agent, with the goal of increased robustness and usability.

References

ALTHOFF, F., MCGLAUN, G., SCHULLER, B.., AND LANG, M. 2001. Using Multimodal Interaction to Navigate in Arbitrary Virtual VRML Worlds. *Proceedings of the 2nd Workshop on Perceptual User Interfaces, University of California, Santa Barbara.*

ATHERTON, P. R. 1981. A Method of Interactive Visualization of CAD Surface Models on a Color Video Display. In *Computer Graphics (Proceedings of ACM SIGGRAPH 81), 15,* 3, ACM, 279–287.

AZUMA, R. 1997. A Survey of Augmented Reality. *Presence 6,* 4, 355–385.

BAUCKHAGE, C., FRITSCH, J., ROHLFING, K.J., WACHSMUTH, S., AND SAGERER, G. 2002. Evaluating Integrated Speech and Image Understanding. *Proceedings of IEEE Int. Conf. on Multimodal Interfaces (ICMI '02),* 9–14.

BILLINGHURST, M., SAVAGE-CARMONA, J., OPPENHEIMER, P. AND EDMOND, C. 1995. The Expert Surgical Assistant: An Intelligent Virtual Environment with Multimodal Input. In Weghorst, S., Sieberg, H.B. and Morgan, K.S. (Eds.), *Proceedings of Medicine Meets Virtual Reality IV,* 590–607.

BOLT, R. A. 1980. Put-That-There: Voice and Gesture at the Graphics Interface. *Computer Graphics (Proceedings of SIGGRAPH 80), 14,* 3, ACM, 262–270.

BOLT, R. A., AND HERRANZ, E. 1992. Two-handed gesture in multi-modal dialog., *Proceedings of the ACM Symposium on User Interface Software and Technology,* ACM Press, 7–14.

BRUMITT, B. L., MEYERS, B., KRUMM, J., KERN, A., AND SHAFER, S. 2000. EasyLiving: Technologies for Intelligent Environments. *Proc. 2nd Intl. Symposium on Handheld and Ubiquitous Computing,.* 12–27.

COHEN, P.R., DALRYMPLE, M., MORAN, D.B, PEREIRA, F.C.N, SULLIVAN, J.W., GARGAN, R.A., SCHLOSSBERG, J.L. AND TYLER, S.W. 1989. Synergistic Use of Direct Manipulation and Natural Language. *Conference on Human Factors in Computing Systems (CHI '89),* ACM, 227–233.

COHEN, P. R., JOHNSTON, M., MCGEE, D., OVIATT, S., PITTMAN, J., SMITH, I., CHEN, L., AND CLOW, J. 1997. QuickSet: Multimodal interaction for distributed applications. *Proceedings of the 5th International Multimedia Conference (Multimedia 97),* ACM Press, 31–40.

COHEN, P. R., MCGEE, D., OVIATT, WU, L., CLOW, J., KING, R., JULIER, S., AND ROSENBLUM, L. 1999. Multimodal Interaction for 2D and 3D Environments. *IEEE Computer Graphics and Applications, 19,* 4, 10–13.

FRÖHLICH, B., PLATE, J., WIND, J., WESCHE, G., AND GÖBEL, M. 2000. Cubic-Mouse-Based Interaction in Virtual Environments. *IEEE Computer Graphics and Applications 20,* 4, 12–15.

HAUPTMANN, A. G. 1989. Speech and Gestures for Graphic Image Manipulation, *Conference on Human Factors in Computing Systems (CHI '89),* ACM, 241–245.

IBA S., PAREDIS. C. J. J., AND KHOSLA P. K. 2002. Interactive Multi-Modal Robot Programming. In *Proceedings of the International Conference on Robotics and Automation,* 161–168.

INSELBERG, A. AND DIMSDALE, B. *1990. Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry. Proceedings of IEEE Visualization 90, 361–378.*

JOHNSTON, M., COHEN, P. R., MCGEE, D. R. OVIATT, S. L., PITTMAN, J. A., AND SMITH, I. 1997. Unification-based multimodal integration. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL/Coling 97).* ACL Press, 281–288.

JOHNSTON, M. 1998. Unification-based multimodal parsing. *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (ACL/Coling 98),* ACL Press, 624–630.

KAISER, C.E. AND COHEN, P. R. 2002. Implementation testing of a hybrid symbolic/statistical multimodal architecture. *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2002),* J. Hansen. Ed., Denver, IEEE Press. 173–176.

KETTEBEKOV, S., YEASIN, M., AND SHARMA, R. 2002. Prosody Based Co-analysis for Continuous Recognition of Coverbal Gestures. *Proceedings of the 4th International Conference on Multimodal Interfaces (ICMI 02)*. IEEE Press. 161–166.

KOONS, D.B., SPARRELL, C.J., AND THÓRISSON, K.R. 1993. Integrating Simultaneous Input from Speech, Gaze, and Hand Gestures. In Maybury, M.T. (ed.), *Intelligent Multimedia Interfaces,* AAAI/MIT Press, 257–276.

KRUM, D.. OMOTESO, O., RIBARSKY, W., STARNER, T., AND HODGES, L. 2002. Speech and Gesture Control of a Whole Earth 3D Visualization Environment. *Proceedings of the Joint Eurographics–IEEE TCVG Symposium on Visualization (VisSym 02)*, 195–200.

LANDRAGIN, F. 2002. The Role of Gesture in Multimodal Referring Actions, Proceedings of the 4th IEEE International Conference on Multimodal Interaction, IEEE Press, Los Alamitos, CA, 173–178.

LATOSCHIK, M. E. 2002. Designing Transition Networks for Multimodal VR-Interactions Using a Markup Language, *Proceedings of the 4th IEEE International Conference on Multimodal Interaction*, IEEE Press, Los Alamitos, CA, 411–416.

LAVIOLA, J. 2000. MSVT: A Virtual Reality-Based Multimodal Scientific Visualization Tool. In *Proceedings of the Third IASTED International Conference on Computer Graphics and Imaging,* November, 1–7.

LIANG, J. AND GREEN, M. 1994. JDCAD: A Highly Interactive 3D Modeling System. *Computers and Graphics 18*, 4, 499–506.

LUCENTE, M., ZWART G-J., AND GEORGE, A.D. 1998. Visualization Space: A testbed for deviceless multimodal user interfaces. *Intelligent Environments '98,* AAAI Spring Symposium Series, 87–92.

MCGEE, D. R., COHEN, P. R., AND OVIATT, S. 1998. Confirmation in Multimodal Systems. *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (ACL/Coling 98)*, ACL Press, 823–829.

NEAL, J. G. AND SHAPIRO, S. C., 1991. Intelligent multimedia interface technology. *Intelligent User Interfaces*, J. W. Sullivan and S. W. Tyler, Eds., ACM Press, 11–43.

OLIVER, N., HORVITZ, E., AND GARG, A. 2002 Layered Representations for Human Activity Recognition, *Proceedings of the 4th IEEE International Conference on Multimodal Interaction*, IEEE Press, Los Alamitos, CA, 3-8.

OVIATT, S. L. 1999. Mutual disambiguation of recognition errors in a multimodal architecture. *Proceedings of the Conference on Human Factors in Computing Systems (CHI '99)*, New York, ACM Press, 576–583.

OVIATT, S. L. 2000. Multimodal Signal Processing in Naturalistic Noisy Environments. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*. 2. B. Yuan, T. Huang and X. Tang , Eds. Beijing. China: Chinese Friendship Publishers. 696–699.

OVIATT, S.L., COHEN, P.R., WU, L.,VERGO, J., DUNCAN, L., SUHM, B., BERS, J., HOLZMAN, T., WINOGRAD, T., LANDAY, J., LARSON, J. AND FERRO, D. 2000. Designing the user interface for multimodal speech and gesture applications: State-of-the-art systems and research directions, Human Computer Interaction, 15(4), 263–322 Reprinted in *Human-Computer Interaction in the New Millennium* (ed. J. Carroll), Addison-Wesley Press, Reading, MA, 2001, 421–456.

PERZANOWSKI, D., SCHULTZ, A., ADAMS, W., AND MARSH, E. 2000. Using a Natural Language and Gesture Interface for Unmanned Vehicles. Unmanned Ground Vehicles II, G.R. Gerhart, R.W. Gunderson, C.M. Shoemaker, Eds., *Aerosense 2000, Proceedings of the Society of Photo-Optical Instrumentation Engineers, 4024*, 341–347.

PODDAR, I., SETHI, Y., OZYILDIZ, AND SHARMA, R.1998. Toward Natural Gesture/Speech HCI: A Case Study of Weather Narration, *Proceedings of the Workshop on Perceptual User Interfaces (PUI 98)*, http://www.cs.ucsb.edu/conferences/PUI/PUIWorkshop98/PUI98.htm.

POUPYREV, I., BILLINGHURST, M., WEGHORST, S., AND ICHIKAWA, T. 1996. Go-Go Interaction Technique: Non-Linear Mapping for Direct Manipulation in VR. *Proceedings of Symposium on User Interface Software and Technology (UIST '96)*. 79–80.

QUEK, F., MCNEILL, D., BRYLL, R., DUNCAN, S., MA, X.-F., KIRBAS, C., MCCULLOUGH, K. E., AND ANSARI, R. 2001. Gesture and speech multimodal conversation. Electrical Engineering and Computer Science Department, University of Illinois, TR VISLab-01-01.

STIEFELHAGEN, R. 2002. Tracking Focus of Attention in Meetings, *Proceedings of the 4th International Conference on Multimodal Interfaces (ICMI 02)*. IEEE Press, 273–380.

TYLER, S. W., SCHLOSSBERG, J. L., GARGAN JR., R. A., COOK, L. K., AND SULLIVAN, J. W. 1991. An intelligent interface architecture for adaptive interaction. *Intelligent User Interfaces*, Sullivan, J. W. and Tyler, S. W. Eds. 45–68.

WAUCHOPE, K. 1994. Eucalyptus: Integrating natural language input with a graphical user interface. Naval Research Laboratory, NRL/FR/5510–94–9711.

WEIMER, D. AND GANAPATHY, K. 1989. A Synthetic Visual Environment with Hand Gesturing and Voice Input. *Proceedings of CHI '89,* ACM Press, 235–240.

WU, L., OVIATT, S., AND COHEN, P.R. 1999. Multimodal integration—A statistical view. *IEEE Transactions on Multimedia, 1*, 4, 334–341.